

Total disc arthroplasty versus anterior cervical interbody fusion: use of the Spine Tango registry to supplement the evidence from randomized control trials

Lukas P. Staub, MD, PhD^a, Christoph Ryser, MD^a, Christoph Röder, MD^a, Anne F. Mannion, PhD^b, Jeffrey G. Jarvik, MD^c, Max Aebi, MD^d, Emin Aghayev, MD^{a,*}

^aInstitute for Evaluative Research in Medicine, Stauffacherstrasse 78, 3014 Bern, Switzerland

^bSpine Centre Division, Department of Teaching, Research and Development, Schulthess Klinik, Lengghalde 2, CH-8008 Zurich, Switzerland

^cComparative Effectiveness, Cost and Outcome Research Centre, University of Washington, 4333 Brooklyn Ave NE, Seattle, WA 98104, USA

^dDepartment of Orthopaedic Surgery, Salem Spital, Schänzlistrasse 39, Bern 3025, Switzerland

Abstract

BACKGROUND CONTEXT: Several randomized controlled trials (RCTs) have compared patient outcomes of anterior (cervical) interbody fusion (AIF) with those of total disc arthroplasty (TDA). Because RCTs have known limitations with regard to their external validity, the comparative effectiveness of the two therapies in daily practice remains unknown.

PURPOSE: This study aimed to compare patient-reported outcomes after TDA versus AIF based on data from an international spine registry.

STUDY DESIGN AND SETTING: A retrospective analysis of registry data was carried out.

PATIENT SAMPLE: Inclusion criteria were degenerative disc or disc herniation of the cervical spine treated by single-level TDA or AIF, no previous surgery, and a Core Outcome Measures Index (COMI) completed at baseline and at least 3 months' follow-up. Overall, 987 patients were identified.

OUTCOME MEASURES: Neck and arm pain relief and COMI score improvement were the outcome measures.

METHODS: Three separate analyses were performed to compare TDA and AIF surgical outcomes: (1) mimicking an RCT setting, with admission criteria typical of those in published RCTs, a 1:1 matched analysis was carried out in 739 patients; (2) an analysis was performed on 248 patients outside the classic RCT spectrum, that is, with one or more typical RCT exclusion criteria; (3) a subgroup analysis of all patients with additional follow-up longer than 2 years (n=149).

RESULTS: Matching resulted in 190 pairs with an average follow-up of 17 months that had no residual significant differences for any patient characteristics. Small but statistically significant differences in outcome were observed in favor of TDA, which are potentially clinically relevant. Subgroup analyses of atypical patients and of patients with longer-term follow-up showed no significant differences in outcome between the treatments.

CONCLUSIONS: The results of this observational study were in accordance with those of the published RCTs, suggesting substantial pain reduction both after AIF and TDA, with slightly greater benefit after arthroplasty. The analysis of atypical patients suggested that, in patients outside the spectrum of clinical trials, both surgical interventions appeared to work to a similar extent to that shown for the cohort in the matched study. Also, in the longer-term perspective, both therapies resulted in similar benefits to the patients. © 2015 Elsevier Inc. All rights reserved.

Keywords:

Propensity score-based matching; Randomized controlled trial; Registry; Spine Tango

FDA device/drug status: Not applicable.

Author disclosures: **LPS:** Nothing to disclose. **CRy:** Nothing to disclose. **CRö:** Nothing to disclose. **AFM:** Nothing to disclose. **JGJ:** Nothing to disclose. **MA:** Nothing to disclose. **EA:** Nothing to disclose.

* Corresponding author. Institute for Evaluative Research in Medicine, Stauffacherstrasse 78, Bern CH-3014, Switzerland. Tel.: +41 31 631 59 30; fax: +41 31 631 59 31.

E-mail address: emin.aghayev@memcenter.unibe.ch (E. Aghayev)

Introduction

The neurologic symptoms and neck pain associated with disc degeneration of the cervical spine can be treated surgically using either the established standard procedure of anterior interbody fusion (AIF) or the more recently introduced total disc arthroplasty (TDA). The evidence regarding the

comparative effectiveness of these surgical procedures has been summarized recently in a Cochrane Review [1]. The review included nine randomized controlled trials (RCTs), five of which were considered to have a low risk of bias [2–6]. It identified evidence for a small but statistically significant difference between the treatments in the alleviation of arm pain, neck-related function, and neurologic outcome, all in favor of arthroplasty [1]. Although statistically significant, the effect sizes were not clinically relevant. Based on these results, the reviewers concluded that at this point in time both treatments could be considered valid therapeutic options. They also cautioned that, as the available RCTs only included follow-up periods of up to 2 years, the hypothesis that TDA results in a reduced incidence of secondary symptoms at adjacent levels would require evaluation in future updates of the review once the results of long-term studies were available.

Randomized control trials have an important limitation in that they often define narrow admission criteria which may not be applicable to daily clinical practice. For instance, some studies only recruited patients younger than 60 years [7,8], limited the treated segments to C3–C7 [2,3,7–12], or excluded marked spondylosis [2,10,11], trauma [5,7,8] or spondylolisthesis, which are sometimes considered contraindications for TDA. Such restrictions are common for RCTs in an effort to examine treatment effects under ideal circumstances. Although these trials provide evidence about the relative *efficacy* of treatments in selected patients, additional studies may be needed that measure the relative *effectiveness* of the treatments in broader populations [13,14]. In doing so, they serve to assess the external validity of the results of RCTs.

The purpose of this study is to demonstrate how data from a surgical registry can be used to supplement the evidence from RCTs with regard to the outcomes after different types of surgery for degenerative cervical spine disease. The Spine Tango registry collects real-world data of patients undergoing surgery of the spine, without the restrictions inherent in RCTs, and some participating clinics follow their patients for a long period of time. Using such prospective cohorts, well designed and carefully executed observational studies can help to answer clinical questions that go beyond the evidence presented by RCTs [15,16]. Spine Tango collects detailed clinical data of patients undergoing TDA and AIF, among other treatments. With the careful application of matching algorithms, a quasi-experimental comparison of the two surgical groups can be achieved to confirm (or refute) RCT results [17]. Registry-based observational studies can also serve to describe the surgical outcome of a broader spectrum of patients outside the narrow admission criteria of RCTs. Finally, the registry gives us the opportunity to perform longer-term effectiveness studies of the surgical treatments.

The aims of this paper were threefold. In a first study we applied the admission criteria used in the available RCTs and conducted a propensity score-adjusted outcome analysis to evaluate whether this design produces short-term results similar to those of the RCTs, for the comparison of TDA and AIF.

We then analyzed the comparative clinical outcome (for TDA vs. AIF) of patients typically excluded from the RCTs published to date, but present in the registry. Third, we evaluated the mid to long-term effects of TDA versus AIF in all patients with follow-up longer than 2 years.

Materials and methods

Spine Tango forms and study cohort

The physician-based surgery form of the registry contains detailed information on diagnosis, the surgical procedure, and surgical and general complications. In addition to these surgical records, patients are asked to complete the self-reported Core Outcome Measures Index (COMI) questionnaire which includes two graphic rating scales (GRS 0–10 points) for neck pain and arm pain [18]. The COMI form is the official patient-reported outcome instrument of the Spine Tango registry [18,19].

None of the centers participating in Spine Tango were involved in any of the RCTs included in the Cochrane review [1]. The registry database was queried in November 2014 for cases with a diagnosis of single-level cervical spine degenerative disc disease, disc herniation, or black disc, treated either by TDA or AIF. The inclusion criteria were as follows: most severely affected segment between C3–C4 and C7/T1/T1, inclusive; no previous surgical treatment of the spine; a baseline COMI; and at least one follow-up COMI completed between 3 months and 2 years postoperatively. If multiple follow-up forms were available for a patient within this time period, the latest form was selected for analysis. Patients who had undergone posterior surgical procedures or whose ASA status was marked as “unknown” in the surgery form were excluded from the analysis.

Outcome measures

The outcome measures were postoperative neck and arm pain and COMI score, as well as neck and arm pain relief and COMI score improvement, with “responders” being defined as those achieving a minimum clinically important change (MCIC) of 2 points in each case [20].

Matching study

For the first study we applied the following exclusion criteria to the Spine Tango cohort to reflect the criteria applied in the published RCTs: age ≥ 60 years, follow-up time > 2 years, treated segment C7–T1, diagnosis of spondylosis, trauma, facet joint degeneration, or spondylolisthesis. Due to incompatibility across different generations of the surgery form, information on spondylosis was missing in 40% and that on facet joint degeneration in 60% of the study cohort. In these cases, we assumed the information to be missing at random and calculated the proportions of spondylosis and of facet joint degeneration based on the available data. The data on the remaining additional exclusion criteria were complete.

With this restricted sample, we performed a TDA-versus-AIF matched analysis to mimic an RCT setting. The propensity score method was used to adjust for confounding, as described in detail by Rosenbaum and Rubin [21]. In brief, an individual's propensity score is defined as the conditional probability of their receiving TDA as opposed to AIF surgery, given the observed covariates (such as age, gender, previous treatment, American Society of Anesthesiologists (ASA) score, etc.). Two patients with the same propensity score have an equal estimated probability of receiving TDA or AIF: if one receives TDA and the other AIF, the exposure allocation can be considered random, conditional on the observed covariates. Therefore, akin to an RCT, there is balance of the covariates between exposure groups after adjusting for the propensity score. The important difference between propensity score adjustment and RCTs is that the latter are able to balance both measured and unmeasured covariates. Propensity scores can only control for the measured covariates.

The individual propensity scores were obtained from a multivariable logistic regression model with the following covariates: baseline age (continuous), gender (male or female), degenerative disc disease (yes or no), disc herniation (yes or no), previous conservative treatment (none, <6 months, 6–12 months, >12months), ASA status (ASA 1, ASA 2, >ASA 2), segment (C3–C4, C4–C5, C5–C6, C6–C7), preoperative neck and arm pain scores (continuous), COMI score (continuous), and length of COMI follow-up in days (continuous). The propensity scores were then fed into a greedy matching algorithm for 1:1 TDA-to-AIF matching, using the “OneToManyMTCH” SAS Macro presented by Parsons [22].

For the continuous study outcomes, mean differences and 95% confidence intervals (CIs) between TDA and AIF groups were calculated. In addition, differences between the treatment groups in the proportion of responders (those achieving the MCIC) were expressed as relative risks (RRs) and 95% CIs and the number needed to treat (NNT) where appropriate.

Atypical patients study

In the second study, an analysis of patients outside the classic RCT spectrum was performed. This study sample comprised all (atypical) patients from the overall cohort that were not used in the matching study because they had one or more of the exclusion criteria defined in the first study. A detailed description of the characteristics of these patients was given to demonstrate the extent to which these patients differed from the RCT patient population.

For the continuous study outcomes, multivariable linear regression models, allowing for significant baseline differences between the treatment groups, were fitted to assess the relative effectiveness of the two types of treatment. Similarly, multivariable logistic regression models, controlling for significant between-group baseline differences, were built to obtain adjusted estimates of the proportion of responders in the two treatment groups, which were expressed as odds ratios (ORs) and 95% CIs and the NNT where appropriate.

Long-term study

Lastly, a subgroup analysis was carried out of all patients with additional longer-term follow-up, that is, at least 2 years postoperatively. As in the atypical patients study, the relative effectiveness of the two types of treatment was assessed using multivariable regression analyses, allowing for significant between-group baseline differences.

The stability of all the multivariable regression models was assessed using the Hosmer and Lemeshow goodness-of-fit test. The significance level was set to 0.05 throughout the study. All statistical analyses were conducted using SAS 9.4 (SAS Institute, Inc., Cary, NC, USA).

Results

Out of a total of more than 75,000 Spine Tango surgery forms in the registry, the selection process for the present study resulted in a cohort of 987 patients undergoing TDA or AIF surgery. Fig. 1 shows the study profile of the included patients. Overall, 35 hospitals from 8 countries (Australia, Belgium, Germany, Poland, Slovenia, Switzerland, UK, and the USA) contributed their data to this study. TDA was documented in 30 hospitals from 8 countries. AIF was documented in 16 hospitals from 5 countries.

Matching study

Out of the 987 patients, 739 cases were identified as being representative of RCT patients. Of these, 190 patients underwent TDA and 549 AIF. All TDA cases found an AIF counterpart, and the matching process resulted in 2 groups with 190 patient pairs with well-balanced patient characteristics (Table 1), leaving 359 AIF patients without a partner.

The outcome analysis in the matched patients produced statistically significant group differences for postoperative neck pain (mean difference [MD] 0.6 points; 95% CI 0.0, 1.2; $p=.04$), arm pain (MD 0.7; 95% CI 0.1, 1.3; $p=.02$), and COMI score (MD 0.8 points; 95% CI 0.2, 1.4; $p=.01$), all in favor of TDA (Table 2; Fig. 2). Furthermore, the COMI change score was significantly different in favor of TDA (MD -1.0 ; 95% CI -1.6 , -0.4 ; $p<.01$). Change scores for neck pain (MD -0.5 ; 95% CI -1.1 , 0.2 ; $p=.16$) and arm pain (MD -0.7 ; 95% CI -1.5 , 0.0 ; $p=.06$) did not differ significantly.

The probability of being a responder (ie, achieving an MCIC score of 2 points) for neck pain relief was not significantly different for TDA (62.1%) compared with AIF (57.9%) (RR 1.07; 95% CI 0.91, 1.26; $p=.40$). The likelihood of being a responder was significantly greater for TDA than for AIF for both arm pain relief (78.4% vs. 67.4%, respectively; RR 1.16; 95% CI 1.03, 1.32; $p=.02$; NNT 10; 95% CI 5, 46) and for COMI score improvement (81.6% vs. 67.9%, respectively; RR 1.20; 95% CI 1.07, 1.35; $p<.01$; NNT 8; 95% CI 5, 20).

Atypical patients study

There were 248 patients who did not meet the common RCT inclusion criteria, representing 25.1% of the overall study

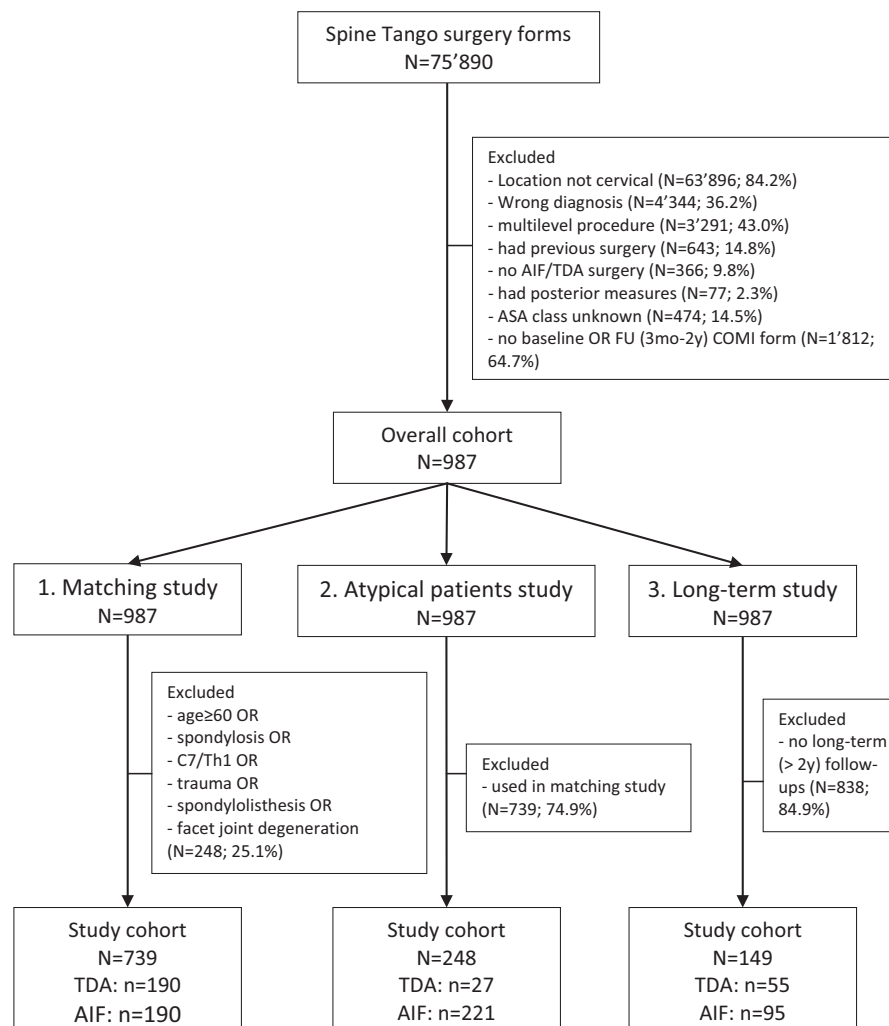


Fig. 1. Study profile of the included patients.

cohort. Table 3 shows that most of these patients were excluded from the matching study because of age ≥ 60 years or a diagnosis of spondylosis. Patients undergoing TDA were generally younger and less likely to have undergone surgery at C7–T1 compared with the AIF patients.

No significant differences in outcome were observed between the two treatment groups (Table 2; Fig. 2). The probability of being a responder (achieving MCIC) with regard to neck pain relief was 59.3% after TDA and 61.5% after AIF. After allowing for patient age, operated segment, and follow-up time, the OR for being a responder after TDA versus AIF was 0.92 (95% CI 0.39, 2.16; $p=.84$). For arm pain relief, the figures were 63.0% and 66.5%, respectively (OR 1.03; 95% CI 0.42, 2.53; $p=.94$), and for COMI, 66.7% and 67.4%, respectively (OR 1.22; 95% CI 0.49, 3.03; $p=.68$).

Longer-term follow-up study

In a total of 149 patients from the overall cohort, an additional longer-term COMI follow-up of more than 2

years after surgery was available. These patients had been treated in seven hospitals in five countries. The mean follow-up time was 55.0 (± 12.2) months (range 27.0–76.5 months).

The characteristics of the patients in the longer-term follow-up study are shown in Table 4. Patients who had received AIF were significantly older than patients in the TDA group. Compared with the AIF group, TDA patients had significantly lower COMI scores and significantly greater COMI score improvements at the short-term follow-up (Table 2; Fig. 2). However, at the longer-term follow-up, the differences between the treatment groups failed to reach significance.

At the longer-term follow-up, the probability of being a responder (achieving MCIC) with regard to neck pain relief was 63.6% for TDA and 64.9% for AIF. After allowing for patient age, the OR for being a responder after TDA versus AIF was 1.02 (95% CI 0.50, 2.11; $p=.95$). For arm pain relief the corresponding figures were 80.0% and 64.9%, respectively (OR 2.25; 95% CI 1.00, 5.09; $p=.05$; NNT 7; 95% CI

Table 1

Matching study: patient characteristics in matched patients (n=380) and non-matched patients (n=359)

Patient characteristics	Matched patients			Non-matched AIF patients (n=359)
	TDA (n=190)	AIF (n=190)	p-Value	
Mean age in years (SD)	44.4 (7.5)	44.2 (7.7)	.71	49.2 (6.8)
Age range	22.5–58.4	23.8–59.9	—	26.9–60.0
Female [%]	53.7	55.3	.76	56.6
DD [%]	24.7	22.1	.54	24.0
DH [%]	92.6	87.9	.12	90.0
No previous treatment [%]	9.5	9.0		21.2
<6 months conservative treatment [%]	60	60.5	1.0	40.7
6–12 months conservative treatment [%]	15.3	15.8		19.2
>12 months conservative treatment [%]	15.3	14.7		18.9
ASA 1 [%]	55.3	65.3		54.6
ASA 2 [%]	42.6	32.6	.13	42.6
ASA>2 [%]	2.1	2.1		2.8
C3/4 [%]	1.1	2.6		2.8
C4/5 [%]	4.2	5.8	.54	8.1
C5/6 [%]	48.4	49.5		44.0
C6/7 [%]	46.3	42.1		45.1
Time from operation to last available follow-up examination [months] (SD)	16.8 (8.1)	16.7 (7.8)	.86	12.7 (8.2)
Neck pain baseline (SD)	5.4 (2.7)	5.6 (2.9)	.55	5.8 (2.8)
Arm pain baseline (SD)	6.2 (2.6)	6.2 (2.8)	.96	6.5 (2.6)
COMI score baseline (SD)	7.5 (1.7)	7.4 (1.8)	.45	7.2 (1.9)

TDA, total disc arthroplasty; AIF, anterior interbody fusion; COMI, Core Outcome Measures Index; SD, standard deviation; DD, degenerative disc, DH, disc herniation; ASA, American Society of Anesthesiologists score.

Note: Paired *t* test and Chi-square test were used for comparisons between the matched pairs, as appropriate.

4, 126), and for COMI, 76.4% and 68.1%, respectively (OR 1.44; 95% CI 0.65, 3.18; *p*=.36).

Discussion

Summary of results

In this analysis of surgical outcomes after TDA and AIF we demonstrate how data from the Spine Tango registry can be used to present evidence that both complements and supplements that provided by RCTs. The first study, a matched comparison of patients who were representative of the type recruited into the RCTs published to date, showed that the outcomes for TDA were slightly but statistically significantly superior to those for AIF. Our second study examined patients that are typically excluded from the RCTs (eg, due to restrictions on age or particular diagnoses) but were documented in Spine Tango. Here, no differences between the surgical groups were observed in the treatment effects. For the third study, we analyzed the surgical outcome of registry patients with longer-term follow-up. Although there were clear trends for superior results with TDA, only borderline statistically significant differences between TDA and AIF were observed for any of the outcomes.

Clinical implications

The matched study confirmed the main results of the existing RCTs that there is a tendency (sometimes statistically significant) for clinical results to be in favor of TDA. The post-operative pain levels and COMI scores were significantly lower

in the TDA group, showing slightly larger differences between the surgical groups than those in the Cochrane review. Although the differences in our study were still relatively small, more than 10% more of the patients in the TDA group achieved the MCIC for arm pain and COMI compared with the AIF group. The RR of being a responder in relation to arm pain (RR=1.16) and COMI score (RR=1.20) was small, with corresponding NNTs of 10 and 8. These NNTs are nonetheless potentially clinically relevant [23], especially in view of the literature reporting no greater harms in TDA than AIF [24]; however, they should be confirmed in future studies in which the relative costs and risks of each treatment (number needed to harm) are directly evaluated. Such studies should also be of a larger size, to obtain tighter confidence intervals around the NNTs.

The study of atypical patients suggests that, in patients outside the spectrum of clinical trials, both surgical interventions appear to work to a similar extent. No superiority of TDA could be observed. Fig. 2 (left column) shows that all patients achieved comparable levels of pain alleviation and COMI score improvement in the first 2 years after surgery.

The analysis of long-term outcomes revealed that the results of surgery remained relatively constant over time. Regardless of the treatment group, the improvements in neck pain, arm pain, and COMI scores persisted throughout the 5 years of observation. Although only borderline statistically significant, arm pain relief showed a tendency for more favorable results with TDA than with AIF, with an NNT of 7. Again, larger studies are needed to confirm our findings with a higher precision.

Table 2

Outcomes measured in the three studies

Matching study (n=739)					
Outcome	TDA (n=190)	AIF (n=190)	Mean difference	95% CI	p-Value
Neck pain last available COMI follow-up (SE)	2.7 (0.2)	3.3 (0.2)	0.6	0.0–1.2	.04
Neck pain change (SE)	2.7 (0.2)	2.3 (0.2)	–0.5	–1.1–0.2	.16
Arm pain last available COMI follow-up (SE)	2.2 (0.2)	2.9 (0.2)	0.7	0.1–1.3	.02
Arm pain change (SE)	4.0 (0.3)	3.3 (0.3)	–0.7	–1.5–0.0	.06
COMI score last available COMI follow-up (SE)	2.8 (0.2)	3.7 (0.2)	0.8	0.2–1.4	.01
COMI score change (SE)	4.7 (0.2)	3.7 (0.2)	–1.0	–1.6–0.4	<.01
Atypical patients study (n=248)					
Outcome	TDA (n=27)	AIF (n=221)	Adjusted mean difference	95% CI	p-Value
Neck pain last available COMI follow-up (SE)	2.6 (0.5)	2.7 (0.2)	0.1	–1.0–1.2	.85
Neck pain change (SE)	2.7 (0.7)	2.6 (0.2)	–0.1	–1.5–1.2	.84
Arm pain last available COMI follow-up (SE)	2.2 (0.6)	2.5 (0.2)	0.3	–0.9–1.5	.65
Arm pain change (SE)	3.4 (0.7)	3.1 (0.3)	–0.3	–1.7–1.2	.73
COMI score last available COMI follow-up (SE)	3.1 (0.6)	3.3 (0.2)	0.2	–0.9–1.4	.70
COMI score change (SE)	3.9 (0.6)	3.6 (0.2)	–0.3	–1.5–0.9	.61
Long-term follow-up study (n=149)					
Outcome	TDA (n=55)	AIF (n=94)	Adjusted mean difference	95% CI	p-Value
Short-term outcome					
Neck pain last available COMI follow-up (SE)	2.1 (0.4)	3.4 (0.3)	1.3	0.4–2.3	<.01
Neck pain change (SE)	3.0 (0.4)	2.2 (0.3)	–0.9	–1.9–0.2	.11
Arm pain last available COMI follow-up (SE)	2.1 (0.4)	3.1 (0.3)	1.0	0.0–2.0	.05
Arm pain change (SE)	3.7 (0.5)	3.0 (0.4)	–0.7	–2.0–0.6	.27
COMI score last available COMI follow-up (SE)	2.2 (0.4)	3.6 (0.3)	1.4	0.4–2.3	<.01
COMI score change (SE)	5.2 (0.4)	3.7 (0.3)	–1.5	–2.6–0.4	<.01
Longer-term outcome					
Neck pain last available COMI follow-up (SE)	2.4 (0.4)	3.2 (0.3)	0.7	–0.2–1.7	.14
Neck pain change (SE)	2.7 (0.4)	2.4 (0.4)	–0.3	–1.3–0.8	.65
Arm pain last available COMI follow-up (SE)	2.0 (0.4)	3.0 (0.3)	1.0	0.0–2.0	.05
Arm pain change (SE)	3.8 (0.5)	3.1 (0.4)	–0.7	–2.0–0.5	.25
COMI score last available COMI follow-up (SE)	2.6 (0.4)	3.5 (0.3)	0.9	–0.1–1.9	.09
COMI score change (SE)	4.8 (0.5)	3.8 (0.3)	–1.0	–2.2–0.1	.08

TDA, total disc arthroplasty; AIF, anterior interbody fusion; COMI, Core Outcome Measures Index; SE, standard error; CI, confidence interval.

Note: Multivariable linear regression models allowing for patient age, operated segment and follow-up time were used to obtain adjusted estimates of mean differences.

Overall, our analyses provide confirmation that AIF is similarly effective in the different populations, whereas TDA seems to work significantly better within the restricted inclusion criteria (possibly reflecting tighter indications for surgery) of RCTs and our matched analysis. Whereas other studies have found a slight but constant decrease of neck pain and arm pain in the 5 years after TDA [25], we did not see any such change; in addition, we found no significant differences between the (smaller) TDA and AIF groups in their long-term outcomes.

Research implications

The results of randomized controlled trials have been criticized for their limited applicability to everyday clinical practice [13,26–28]. Several factors can affect the external validity of the results of any study. First, the study setting determines

the patients to which the study results will apply. One important issue here is the definition of eligibility criteria. For example, investigators may decide to only recruit young patients who are likely to respond well to spinal surgery. Pre-randomization ineligibility of up to 90% has been reported due to narrow inclusion criteria of clinical trials [29], which has the potential to severely restrict the generalizability of trial results. In our cohort the situation was less accentuated: about three quarters of the patients in the Spine Tango registry were comparable to the patients in the available trials, and only the remaining quarter had one or more of the “typical” exclusion criteria used in the RCTs. It appears that the available RCTs used a pragmatic (rather than a tightly controlled explanatory) approach [30] in that they did not overly restrict their study samples but strove to measure treatment effectiveness in a setting similar to clinical practice. Nevertheless, because of the complete absence of inclusion criteria

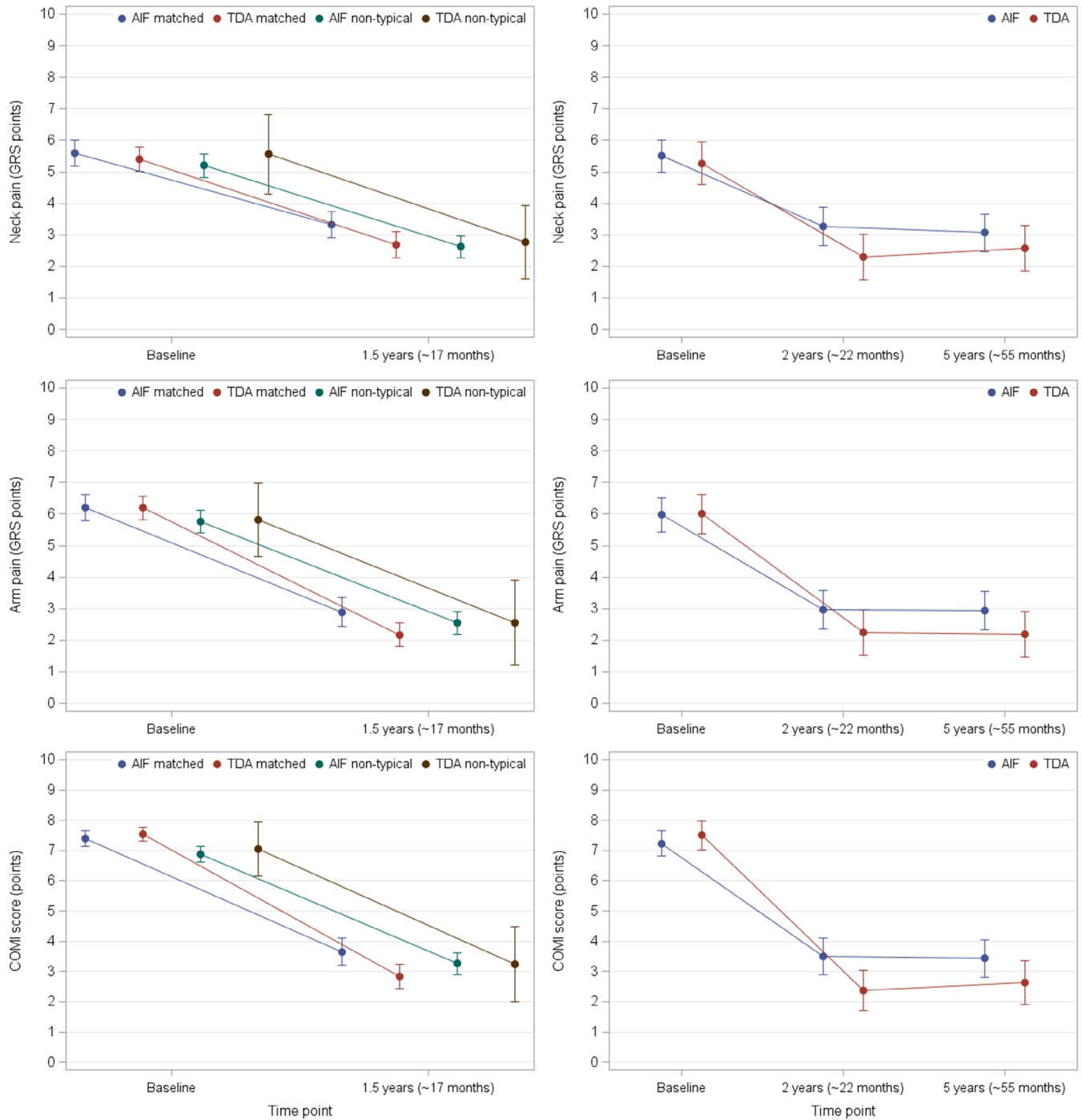


Fig. 2. Neck and arm pain and COMI scores at baseline and short-term follow-up (Left) and longer-term follow-up (Right). Unadjusted estimates and standard deviations are shown.

in the registry we were able to analyze 248 patients outside the spectrum of the trials. The documentation of these atypical patients is one of the major strengths of clinical registries.

The second question affecting the external validity of trial results is whether the outcome measured is relevant in clinical practice. For practical reasons, investigators often choose to only measure the short-term efficacy of treatments. But an initial response to surgery is not necessarily a good predic-

tor of long-term benefit. The authors of the Cochrane review considered it an important limitation that the available RCTs comparing TDA and AIF followed their patients for only 2 years, and they expressed the hope that evidence of the longer-term outcome would be available for future updates of the review. In view of the high costs associated with long-term follow-up of patients within a controlled trial, it may be more efficient to undertake well-designed observational studies at

Table 3

Atypical sample study: patient characteristics and reasons for exclusion from matching study (n=248)

Patient characteristics	TDA (n=27)	AIF (n=221)	p-Value
Mean age in years (SD)	53.8 (12.8)	61.1 (11.5)	.003
Age range	27.3–76.3	26.4–88.7	—
Female [%]	33.3	48.4	.14
DD [%]	51.9	39.8	.23
DH [%]	70.4	78.7	.32
No previous treatment [%]	11.1	14.0	
<6 months conservative treatment [%]	33.3	40.7	
6–12 months conservative treatment [%]	18.5	22.2	.47
>12 months conservative treatment [%]	37.0	23.1	
ASA 1 [%]	40.7	27.6	
ASA 2 [%]	55.6	61.1	.24
ASA>2 [%]	3.7	11.3	
C3/4 [%]	—	10.0	
C4/5 [%]	7.4	11.8	
C5/6 [%]	63.0	33.5	.037
C6/7 [%]	18.5	24.4	
C7/T1 [%]	11.1	20.4	
Time from operation to last available follow-up examination [months] (SD)	17.5 (7.5)	14.2 (8.0)	.033
Neck pain baseline (SD)	5.6 (3.2)	5.2 (2.9)	.52
Arm pain baseline (SD)	5.8 (2.9)	5.8 (2.7)	.88
COMI score baseline (SD)	7.1 (2.3)	6.9 (2.0)	.47
Reasons for exclusion from matching study			
Age≥60 years [%]	40.7	65.6	.012
Spondylosis [%]	50.0	41.9	.50
C7–T1 [%]	11.1	20.4	0.25
Spondylolisthesis [%]	3.7	3.2	1.0
Facet joint arthritis [%]	—	11.5	1.0
Trauma [%]	3.7	—	0.11

TDA, total disc arthroplasty; AIF, anterior interbody fusion; COMI, Core Outcome Measures Index; SD, standard deviation; DD, degenerative disc; DH, disc herniation; ASA, American Society of Anesthesiologists score.

Note: Wilcoxon rank-sum test and Chi-square test were used for comparisons between the treatment groups, as appropriate. Statistically significant differences are in bold.

this stage. There is no controlled allocation of patients to treatment groups in observational studies, which simplifies their design and administration and reduces costs. Interestingly, the lack of randomization does not seem to generally affect the estimates of treatment effectiveness. As also seen in our first study, RCTs and observational studies often report similar results: Benson and Hartz reviewed 83 RCTs and 53 observational studies and found that in only 2 out of 19 therapeutic comparisons did the estimates of the treatment effects in observational studies lie outside the 95% CI for the combined effect in the RCTs [31]. Concato et al. reviewed the medical literature over a 5-year period and concluded that the results of well-designed observational studies with either a cohort or a case-control design did not systematically overestimate the magnitude of the effect of treatment as compared with that based on RCTs on the same topic [32]. Bhandari et al. showed that, when adjusting for important risk factors, the results of observational studies (n=13) on revision and mortality rates after internal fixation of femoral neck fracture were similar to those from RCTs (n=14) [33]. Previous registry studies have also shown similar results to RCTs for the treatments examined in the present study [34] as well as for other spine surgical treatments such as balloon kyphoplasty [35].

Finally, the success of new treatments depends on how safe and acceptable they prove to be in practice. Again, registries are well suited to answer these research questions. They allow independent high-quality observational studies in routine care. If the appropriate statistical models are used to estimate treatment effectiveness (eg, propensity score adjustment), these studies can help answer the important questions about what happens when RCT results are applied to real-life clinical practice.

Limitations

Spine Tango is a voluntary registry, making it potentially prone to selection bias at several levels. The participating clinics are not necessarily representative of all spine surgical centers in their countries. However, our study was based on data from 35 hospitals in 8 countries, including a range from small regional hospitals to university clinics and large spine centers. There was no control as to whether the participating clinics documented all their patients in Spine Tango or whether a selection of cases occurred. Further, within documented patients, we cannot know whether incomplete follow-up reporting occurred at random or not. These challenges

Table 4
Longer-term follow-up study: patient characteristics (n=149)

Patient characteristics	TDA (n=55)	AIF (n=94)	p-Value
Mean age in years (SD)	44.3 (8.7)	50.6 (10.9)	.001
Age range	22.5–62.6	24.1–81.7	—
Female [%]	50.9	56.4	.52
DD [%]	38.2	37.2	.91
DH [%]	80.0	79.8	.98
No previous treatment [%]	5.5	8.5	.81
<6 months conservative treatment [%]	50.9	44.7	
6–12 months conservative treatment [%]	18.2	17.0	
>12 months conservative treatment [%]	25.5	29.8	
ASA 1 [%]	61.8	51.1	.43
ASA 2 [%]	36.4	45.7	
ASA>2 [%]	1.8	3.2	
C3/4 [%]	—	3.2	.23
C4/5 [%]	3.6	6.4	
C5/6 [%]	56.4	44.7	
C6/7 [%]	40	41.5	
C7/T1 [%]	—	4.3	
Neck pain baseline (SD)	5.3 (2.5)	5.5 (2.4)	.53
Arm pain baseline (SD)	6.0 (2.3)	6.0 (2.7)	.88
COMI score baseline (SD)	7.5 (1.8)	7.2 (2.1)	.55

TDA, total disc arthroplasty; AIF, anterior interbody fusion; COMI, Core Outcome Measures Index; SD, standard deviation; DD, degenerative disc; DH, disc herniation; ASA, American Society of Anesthesiologists score.

Note: Wilcoxon rank-sum test and Chi-square test were used for comparisons between the treatment groups, as appropriate. Statistically significant differences are in bold.

apply to many studies (including some controlled trials) and can probably only be met by audits and other monitoring tools that ensure the quality of documentation for all participants.

Expectation bias is another potential problem of observational registry data collected in an open-label fashion or indeed of non-blinded RCTs. Patients may be biased in their reporting of outcome when they know what surgical procedure they have received, with more favorable results being expected for the “novel” treatment arm. However, in contrast to an RCT—in which patients must be explicitly informed about the nature of the trial and the competing treatments under investigation—patients in a registry may not be aware of having received a “less sophisticated” alternative, or the latter may not even have been an option at the time of their treatment.

Finally, observational studies are not able to control for all potential confounding variables. Even with sophisticated analytical methods such as propensity score adjustment, there is always a risk of bias due to unmeasured covariates.

Way forward

No single study design is able to comprehensively assess a given surgical treatment through all its stages of evaluation, from initial efficacy testing to the establishment of real-world effectiveness and cost-effectiveness compared with other treatment options. In 1996 Black stated, “randomized trials provide an indication of the minimum effect of an intervention whereas observational studies offer an estimate of the

maximum effect. If this is the case, policymakers need data from both approaches when making decisions about health services, and neither should reign supreme” [14]. More recently Wouter and colleagues suggested that both prospective controlled trials and postmarketing registries are necessary to demonstrate the safety and effectiveness of new implant devices [36]. We share these views. The randomized controlled trial is the established “gold standard” of study designs to answer the question as to whether a new treatment can work in a controlled setting. Ideally, these RCTs demonstrate whether the new treatment works better than (or is at least equal to) the current standard treatment modality, rather than using placebo as a comparator. Once the superiority (or non-inferiority) of the new treatment has been established, further studies may be needed to assess the effectiveness of the new treatment for patients in whom it was not initially examined [37].

Consistent and sufficiently detailed data capture in registries like Spine Tango can give us important knowledge about new treatments and their most effective and safe application [25,35,38]. Mandatory high coverage registries such as Swespine [39] or SWISSpine [25,35] have provided solid data about new treatments and present key information to policymakers, supplementary to that delivered by RCTs. From the process point of view, a well-integrated documentation system, which many of the Spine Tango participants have achieved, may be less cumbersome for the clinical workflow than an on-and-off data collection in clinical studies. The need for cost-effective, multi-sourced, and widely shareable data collection has never been greater [40]. The interest in registries is constantly growing, and rightly so.

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.spinee.2015.11.056.

References

- [1] Boselie TF, Willems PC, van Mameren H, de Bie R, Benzel EC, van Santbrink H. Arthroplasty versus fusion in single-level cervical degenerative disc disease. *Cochrane Database Syst Rev* 2012;(9): CD009173.
- [2] Heller JG, Sasso RC, Papadopoulos SM, et al. Comparison of BRYAN cervical disc arthroplasty with anterior cervical decompression and fusion: clinical and radiographic results of a randomized, controlled, clinical trial. *Spine* 2009;34:101–7.
- [3] Mummaneni PV, Burkus JK, Haid RW, Traynelis VC, Zdeblick TA. Clinical and radiographic analysis of cervical disc arthroplasty compared with allograft fusion: a randomized controlled clinical trial. *J Neurosurg Spine* 2007;6:198–209.
- [4] Murzluff J, McConnell J, Tomaras C, Peppelman W, Volcan I, Baker K. 2-Year multicenter follow-up in a prospective randomized clinical trial: comparison of a cervical artificial disc to an ACDF treatment. *Spine J* 2010;10(9 Suppl. 1):135S–6S.
- [5] Phillips FM, Lee JY, Geisler FH, et al. A prospective, randomized, controlled clinical investigation comparing PCM cervical disc arthroplasty with anterior cervical discectomy and fusion. 2-year results from the US FDA IDE clinical trial. *Spine* 2013;38:E907–18.
- [6] Kelly MP, Mok JM, Frisch RF, Tay BK. Adjacent segment motion after anterior cervical discectomy and fusion versus ProDisc-C cervical total disc arthroplasty: analysis from a randomized, controlled trial. *Spine* 2011;36:1171–9.
- [7] Murrey D, Janssen M, Delamarter R, et al. Results of the prospective, randomized, controlled multicenter Food and Drug Administration investigational device exemption study of the ProDisc-C total disc replacement versus anterior discectomy and fusion for the treatment of 1-level symptomatic cervical disc disease. *Spine J* 2009;9:275–86.
- [8] Nabhan A, Ahlhelm F, Shariat K, et al. The ProDisc-C prosthesis: clinical and radiological experience 1 year after surgery. *Spine* 2007;32:1935–41.
- [9] Davis RJ, Kim KD, Hisey MS, et al. Cervical total disc replacement with the Mobi-C cervical artificial disc compared with anterior discectomy and fusion for treatment of 2-level symptomatic degenerative disc disease: a prospective, randomized, controlled multicenter clinical trial: clinical article. *J Neurosurg Spine* 2013;19:532–45.
- [10] Garrido BJ, Taha TA, Sasso RC. Clinical outcomes of Bryan cervical disc arthroplasty: a prospective, randomized, controlled, single site trial with 48-month follow-up. *J Spinal Disord Tech* 2010;23:367–71.
- [11] Sasso RC, Smucker JD, Hacker RJ, Heller JG. Artificial disc versus fusion: a prospective, randomized study with 2-year follow-up on 99 patients. *Spine* 2007;32:2933–40, discussion 41–2.
- [12] Coric D, Cassis J, Carew JD, Boltes MO. Prospective study of cervical arthroplasty in 98 patients involved in 1 of 3 separate investigational device exemption studies from a single investigational site with a minimum 2-year follow-up. Clinical article. *J Neurosurg Spine* 2010;13:715–21.
- [13] Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” *Lancet* 2005;365:82–93.
- [14] Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215–18.
- [15] Castillo RC, Scharfstein DO, MacKenzie EJ. Observational studies in the era of randomized trials: finding the balance. *J Bone Joint Surg Am* 2012;94(Suppl. 1):112–17.
- [16] Hoppe DJ, Schemitsch EH, Morshed S, Tornetta P 3rd, Bhandari M. Hierarchy of evidence: where observational studies fit in and why we need them. *J Bone Joint Surg Am* 2009;91(Suppl. 3):2–9.
- [17] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399–424.
- [18] Mannion AF, Porchet F, Kleinstuck FS, et al. The quality of spine surgery from the patient’s perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *Eur Spine J* 2009;18(Suppl. 3):374–9.
- [19] Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. *Spine* 1998;23:2003–13.
- [20] Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine* 2008;33:90–4.
- [21] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for casual effects. *Bimetrika* 1983;70:41–55.
- [22] Parsons LS. Performing a 1:N case-control match on propensity score. Available at: <http://www2.sas.com/proceedings/sugi29/165-29.pdf>. Accessed November 7, 2013.
- [23] Citrome L. Compelling or irrelevant? Using number needed to treat can help decide. *Acta Psychiatr Scand* 2008;117:412–19.
- [24] Anderson PA, Hashimoto R. Total disc replacement in the cervical spine: a systematic review evaluating long-term safety. *Evid Based Spine Care J* 2012;3(S1):9–18.
- [25] Aghayev E, Barlocher C, Sgier F, et al. Five-year results of cervical disc prostheses in the SWISSspine registry. *Eur Spine J* 2013;22:1723–30.
- [26] Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials* 2006;1:e9.
- [27] van der Windt D, Croft P. From the particular to the universal—how does an efficacy trial translate to practice? *Pain* 2011;152:967–8.
- [28] Ahmad N, Boutron I, Moher D, Pitrou I, Roy C, Ravaut P. Neglected external validity in reports of randomized trials: the example of hip and knee osteoarthritis. *Arthritis Rheum* 2009;61:361–9.
- [29] Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *Br Med J (Clin Res Ed)* 1984;289:1281–4.
- [30] Roland M, Torgerson DJ. What are pragmatic trials? *BMJ* 1998;316:285.
- [31] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–86.
- [32] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–92.
- [33] Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the *Journal of Bone and Joint Surgery* from 1988 through 2000. *J Bone Joint Surg Am* 2002;84-A:388–96.
- [34] Grob D, Porchet F, Kleinstuck FS, et al. A comparison of outcomes of cervical disc arthroplasty and fusion in everyday clinical practice: surgical and methodological aspects. *Eur Spine J* 2010;19:297–306.
- [35] Hubschle L, Borgström F, Olafsson G, et al. Real-life results of balloon kyphoplasty for vertebral compression fractures from the SWISSspine registry. *Spine J* 2014;14:2063–77.
- [36] Moojen WA, Bredenoord AL, Viergever RF, Peul WC. Scientific evaluation of spinal implants: an ethical necessity. *Spine* 2014;39:2115–18.
- [37] Jarvinen TL, Sievanen H, Kannus P, Jokihaara J, Khan KM. The true cost of pharmacological disease prevention. *BMJ* 2011;342:d2175.
- [38] Munting E, Röder C, Sobottke R, Dietrich D, Aghayev E. Patient outcomes after laminotomy, hemilaminectomy, laminectomy and laminectomy with instrumented fusion for spinal canal stenosis: a propensity score based study from the Spine Tango registry. *Eur Spine J* 2014;24:358–68.
- [39] Stromqvist B, Fritzell P, Hagg O, Jonsson B, Sanden B. Swedish Society of Spinal S. Swespine: the Swedish spine register: the 2012 report. *Eur Spine J* 2013;22:953–74.
- [40] Mandl KD, Edge S, Malone C, Marsolo K, Natter MD. Next-generation registries: fusion of data for care, and research. *AMIA Summits Transl Sci Proc* 2013;2013:164–7.